Reducing Preventable Service Visits With Generative AI: Altice USA & Palantir

A technical paper prepared for presentation at SCTE TechExpo24

Gavin Mitchell
VP, Product Quality Assurance
Altice USA
Gavin.Mitchell@alticeusa.com

Alex Gottwald
Head of Telecom, North America
Palantir Technologies
agottwald@palantir.com

Austin Atmaja, Palantir Technologies

Bruce Gatete, Palantir Technologies

Shane McWilliams, Palantir Technologies

Kate Van Horn, Palantir Technologies

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

1. Introduction

Preventable service visits pose a significant burden to operators in the form of financial costs to the business, misallocated technician time, and degraded customer experience. Altice USA and Palantir set out to take a user and data-driven approach to solving this problem, leveraging the latest advancements in machine learning, generative AI, and data modeling.

Altice USA and Palantir built a network and customer model, replicating real-world business workflows in a digital twin. This model utilizes data from an array of sources including physical network topology for both Hybrid Fiber-Coax (HFC) as well as Fiber-to-the-Home (FTTH) networks, customer and device-based service telemetry, customer contact points and technician service visits. Within the Palantir platform, this data is then leveraged by GenAI and LLM tools to elucidate trends in flows, which allows human operators to enhance their analysis and strategies.

Within 3 months, the initiative team brought new troubleshooting models to production by integrating recommendations within customer care tools indicating to the care operator as to whether a service visit was necessary based on the recent experiences for the customer.

In this paper we will walk through the data collection, modeling and implementation work completed by the team to take the ideas from conception to production. We will also share some of the early findings from the A/B testing which have identified a 7% reduction of preventable service truck rolls compared to a control group and 8% reduction in Average Handle Time (AHT) for care agents. Both represent an improvement in the operations of the business, the experience for our customers and employees. Next steps for further developing models and integration into operational processes will also be discussed.

2. Data Sources

Starting this project, we understood that the quality of the outcome was going to be highly dependent on the strength of the data that we were able to collect and provide into the generative AI models. As we will discuss further in the result and next steps section this is still an area where significant improvement is possible. To start, however, we looked at the universe of information available that was going to provide insight in three categories:

1. Was the information contributing to our understanding of the customer experience for their broadband or video services.
2. Was the information contributing to our understanding of challenges the customer may be having due to issues in the network outside their home.
3. Was the information contributing to our understanding of challenges the customer maybe having inside their home.

In the sections below we have identified the data sets that were included as inputs into our models as well as the frequency with which we were collecting that information.

2.1. Access Networks

The Altice USA network is a mix of HFC as well as FTTH networks. To support the network modeling we looked at telemetry available for both. The table below outlines the data that was used to generate the network model for each.

Table 1 – Access Network Telemetry Included

Dataset Name	Data Update Frequency	Metric Name
Modem RF Signals	Hourly	Signal-to-Noise Ratio (SNR)
		Codeword Error Ratio (CER)
		Received Power
		Transmitted Power
		T3 Timeouts
		T4 Timeouts
Modem Offline Events	Every 15 minutes	Online / Offline Status
Node Health Score	Daily (aggregated)	Node Health Scoring
Node Congestion	Daily (aggregated)	Node Congestion
Network Outage Tickets	N/A	Network Outages

2.2. Broadband CPE

Altice USA supports CPE across DOCSIS® 2.0 through DOCSIS® 3.1 specifications on the HFC network as well as both GPON and XGSPON devices. Most of our devices are integrated gateways comprised of an embedded cable/fiber modem as well as the routing and WIFI components.

Table 2 – Broadband CPE Telemetry

Dataset Name	Data Update Frequency	Metric Name
CPE Metrics	Daily (polled)	Crash count
		CPU Utilization
		Memory Utilization
		Temperature
		Reboot Count

2.3. WIFI Networks

The Altice USA broadband gateways deployed today support a variety of WIFI standards including WIFI 4, WIFI 5, WIFI 6 and WIFI 6E. The table below identifies the WIFI telemetry that was included in the model.

Table 3 – WIFI Network Telemetry

Dataset Name	Data Update Frequency	Metric Name
Wi-Fi Metrics	Daily (aggregated)	Transmit Opportunity
		Backhaul Received Signal Strength Indicator (RSSI)
		Wi-Fi Quality of Experience (QoE) - derived metric that aggregates RSSI values across all clients

2.4. Customer Contacts

History of customer interactions can be indicative of future interactions. To help improve the model we included past customer interactions as part of our model. Where applicable this also included identification of resolutions to past customer problems used for model training.

2.5. Technician Contacts

History of customer service visits interactions can be indicative of future interactions. To help improve the model we included past service visits as part of our model. Where applicable this also included identification of resolutions to past customer problems used for model training.

3. Data Ingest, Normalization, Pipeline Management, and Ontology

In this section we will cover the details surrounding how the above data sources were transitioned from data elements into an overall network and customer model ("*ontology*"). Each section will follow a consistent format: Overview, Technologies Leveraged, Implementation Challenges and Outcomes.

3.1. Data Connection

3.1.1. Overview

The diversity and volume of data sources in the telecom industry, where current IT infrastructures are the product of decades of mergers and acquisitions, present unique challenges, such as ensuring data consistency, security, and real-time availability. When Altice USA and Palantir partnered to employ Palantir software, the initial phase in the data ingest process involved establishing connections to various data sources. For Altice USA, this necessitated the integration of data from multiple systems, including Altice USA's existing technician troubleshooting tool and backend, an existing BigQuery instance, and other relevant sources. This foundational step was critical for creating a unified data environment where machine learning and generative AI models could be effectively applied.

3.1.2. Technologies Leveraged

Within the overall software that was developed to address preventable service visits, two major technologies were leveraged: Palantir's out-of-the-box data connectors and secure data access mechanisms.

Out-of-the-Box Palantir Connectors: Designed to optimize existing software connections and dependencies, these connectors accelerate data integration. By implementing pre-built connectors (both industry-agnostic and telecom-specific) Altice USA was able to streamline the integration process and capture data from in-house tools and enterprise data warehouses. This enabled the capture of customer interactions and service requests, diagnostic data from customer premise equipment (CPE), as well as documentation of historical technician troubleshooting service logs. These pre-built connectors were designed to handle the specific data formats and protocols used by these systems, ensuring smooth and efficient data ingestion in days, rather than weeks. This approach significantly reduced the complexity and time required to integrate diverse data sources into a unified data environment, facilitating faster access to comprehensive datasets for analysis.

Secure Data Access Mechanisms: Given the sensitivity of customer information and critical network data, secure data access was a top priority. Altice USA implemented robust authentication and authorization mechanisms to ensure that only authorized personnel could access sensitive data. This included the use of role-based access control (RBAC), which assigned permissions based on user roles and responsibilities, ensuring that each user had appropriate access levels. Additionally, data encryption was employed both in transit and at rest to protect data from unauthorized access and breaches. Compliance with industry standards and regulations was continuously monitored and enforced to maintain data integrity and confidentiality. This comprehensive security framework ensured that Altice USA could handle sensitive customer data responsibly and securely, fostering trust and compliance.

3.1.3. Implementation and Challenges

The process of establishing data connections involved several key steps and challenges. The identification of data sources and determination of real-time and near-real time data availability were the key steps taken in implementation, while monitoring the challenge of maintaining data consistency and quality.

Data Source Identification: The first step was to identify all relevant data sources within Altice USA's ecosystem. This involved a comprehensive audit of existing systems and other operational data repositories. Each data source was evaluated for its relevance and potential contribution to the machine learning models.

Real-Time & Near Real-Time Data Availability: For the machine learning models to be effective, near real-time customer premise data availability was crucial. This required architecting data pipelines that could retrieve and process data continuously, as soon as the predictive model was invoked from the troubleshooting tool.

Data Consistency and Quality: A major challenge was ensuring data consistency and quality across diverse sources. Data from different systems often had varying formats, structures, and levels of completeness, necessitating a thorough standardization process. Standardization protocols were established to normalize the data, converting it into a common format that could be effectively used in subsequent analysis and modeling.

3.1.4. Outcomes

The successful integration of diverse data sources laid the groundwork for the subsequent steps of data transformation, pipeline management, and ontology creation. By establishing robust data connections, Altice USA was able to create a unified data environment that supported the rapid deployment of machine learning models.

3.2. Data Transformation

3.2.1. Overview

To ensure consistency and usability within data connections, data is transformed through a process of cleaning and formatting. For Altice USA, this step was crucial in preparing the data for effective analysis and model training. The data transformation process involved identifying and rectifying errors, inconsistencies, and missing values, as well as transforming the data into a standardized format suitable for machine learning applications, as represented in the Altice USA ontology. This section details the technologies leveraged and the specific steps taken to achieve comprehensive data transformation.

3.2.2. Technologies Leveraged

Data Cleaning: Altice USA employed tools to identify and rectify errors, inconsistencies, and missing values in data.

Data Formatting: The data was transformed into the desired format through a series of processes, including normalization, ensuring consistency and removing redundancies; aggregation, combining data from multiple sources; and enrichment, enhancing the data by identifying potential connections. These transformations were tailored to the specific needs of the preventable truck roll problem statement and were the underpinnings of the Altice USA network ontology.

3.2.3. Implementation and Challenges

Within the larger processes of data cleaning and data formatting, Altice USA and Palantir focused on error identification and rectification, aggregation and enrichment, and quality assurance and validation.

The data cleaning process began with error identification and rectification. Initially, diagnostic checks were run to detect anomalies such as missing values, duplicate records, and outliers. Once identified, these errors were rectified through automated and manual processes. For instance, missing values were imputed using statistical methods and duplicate records were removed to ensure data integrity.

To enhance the usability of the data, aggregation and enrichment processes were applied. Aggregation involved summarizing data at different levels, such as aggregating network performance metrics by time intervals or geographic regions. Enrichment involved adding additional context to the data, such as appending geographic information to service logs. These processes provided a richer and more comprehensive dataset for analysis.

Finally, ensuring the quality of transformed data was a critical step. Quality assurance checks were implemented to validate the accuracy and consistency of the transformed data. This involved running validation scripts to compare the transformed data against predefined quality metrics and thresholds. Any discrepancies were flagged and addressed to ensure that the data met the required standards.

3.2.4. Outcomes

The successful transformation of data was a pivotal step in preparing the data for machine learning and generative AI applications. By ensuring that the data was accurate, consistent, and enriched with additional context, Altice USA was able to derive high-quality insights from the data. This, in turn, enabled the development of advanced troubleshooting models that could identify patterns and anomalies indicative of potential service issues.

3.3. Pipeline Management

3.3.1. Overview

Effective management of data pipelines is critical for maintaining data integrity and ensuring smooth operations. For Altice USA, managing data pipelines involved monitoring and controlling data workflows to handle the high volumes of data generated from various sources. This step was essential to efficiently and securely process data, enabling the timely deployment of machine learning models and generative AI applications. This section details the technologies leveraged and the specific steps taken to achieve comprehensive pipeline management.

3.3.2. Technologies Leveraged

Health Checks: Automated monitoring tools were employed to continuously assess the performance and health of data pipelines. These tools provided real-time alerts for any issues or anomalies detected in the ingest process, allowing for prompt resolution before causing customer impact.

Permissions Management: Granular control mechanisms were implemented to manage access and permissions. This ensured that only authorized users could modify or access data pipelines, maintaining data security and integrity.

Version Control: Version control systems were used to track changes and maintain versions of data pipelines. This facilitated auditability and rollback capabilities, ensuring that any modifications could be traced and reverted if necessary.

2.2.1 Implementation and Challenges

As the volume of data increased, optimizing and scaling data pipelines became a priority. This involved fine-tuning pipeline configurations to improve processing efficiency and implementing scalable architectures to handle growing data volumes. Load balancing techniques were employed to distribute data processing tasks across multiple nodes, ensuring that pipelines could handle peak loads without performance degradation.

2.2.2 Outcomes

Effective management of data pipelines guarantees data integrity and smooth operations. Automated monitoring and health checks enabled Altice USA to proactively detect and resolve issues, minimizing disruptions. Granular access controls ensured that only authorized personnel could access the data pipelines. Version control systems provided auditability and rollback capabilities, maintaining stability and integrity in a dynamic environment. This framework allowed for efficient management of changes to data workflows.

Optimization and scaling efforts ensured that data pipelines could handle increasing volumes without performance degradation. This was vital for the timely deployment of machine learning models and generative AI applications, which depended on efficient and reliable data processing.

3.4. Ontology Creation

3.4.1. Overview

To make data more accessible and understandable, the creation of an ontology—a structured representation of human-understandable concepts—was a critical step for Altice USA. This process involved defining and mapping key concepts, relationships, and hierarchies within the network data. By creating a comprehensive ontology, Altice USA was able to derive meaningful insights from complex data sets, facilitating more effective analysis and decision-making. This section details the technologies leveraged and the specific steps taken to achieve comprehensive ontology creation.

3.4.2. Technologies Leveraged

Concept Mapping: Tools were used to define and map concepts, relationships, and hierarchies within the data at Altice USA. This included representing objects such as customers, service visits, technicians, nodes, households, care interactions, and troubleshooting tickets. These mappings provided a structured framework for understanding the data.

Semantic Enrichment: The data was enhanced with semantic information to improve searchability and context. This involved adding metadata and annotations to the data, making it easier to query and analyze.

Collaboration: Creating an ontology allowed individuals from previously unconnected parts of the organization. This enabled domain experts to refine and expand one comprehensive ontology. This collaborative approach ensured that the data structure accurately reflected the business context.

3.4.3. Implementation and Challenges

The development of an ontology requires several steps to fully create a digital twin of business operations. Key concepts and relationships are identified and mapped to hierarchies while the data itself undergoes semantic enrichment, collaborative refinement, and validation and iteration processes.

The first step of building out the ontology was to identify the key concepts to be represented. For Altice USA, this included objects such as customers, service visits, technicians, nodes, households, care interactions, and troubleshooting tickets. Each concept was defined in terms of its attributes and relationships with other concepts. For example, a service visit might be linked to a customer, a technician, and a troubleshooting ticket.

Once key concepts and relationships were identified, users leveraged the Palantir Ontology Manager application to map these concepts to broader hierarchies. These mappings provided a clear and organized framework for understanding the data.

While the data was being mapped, semantic enrichment techniques were applied in tandem to enhance the usability of the data. This involved adding metadata and annotations to the data, providing additional context and making it easier to query and analyze. For example, customer data might be enriched with geographic information, and service visit data might be annotated with details about the issues addressed and the outcomes. These enrichments improved the searchability and context of the data, facilitating a more wholistic understanding of the problem statement.

Ensuring the accuracy and relevance of the ontologies was a critical step. Validation processes were implemented to verify that the ontologies accurately represented the data and business context. This involved running test queries and analyses to ensure that the ontologies provided meaningful and accurate results. The ontologies were iteratively refined based on the validation results, ensuring that they remained relevant and accurate.

3.4.4. Outcomes

The successful creation of a comprehensive ontology was a pivotal step in making the data more accessible and understandable. By defining and mapping key concepts, relationships, and hierarchies, Altice USA was able to create a structured framework for understanding the data. This facilitated more effective analysis and decision-making, enabling the identification of patterns and insights that were previously difficult to discern. Importantly, this ontology has already provided a launchpad for additional use cases across other parts of Altice USA: field operations, network, supply chain, customer care, and more.

4. Building The Model

4.1. Feature Engineering

We trained our machine learning models on a comprehensive dataset consisting of customer trouble calls. For each call, we extracted and engineered the following features:

- **Ground Truth Label:** This label indicates whether the problem was inside or outside the home. It was derived based on the outcome of the call. If the call resulted in the technician fixing outside wiring or network issues, the problem was classified as outside the home; otherwise, it was classified as inside the home.

- **Modem RF Signals:** These features measure the quality of the connection between a customer's cable modem and the Cable Modem Termination System (CMTS) over the Hybrid Fiber-Coaxial (HFC) network. Specifically, we considered both upstream and downstream Signal-to-Noise Ratio (SNR), received power (RxPower), transmitted power (TxPower), Codeword Error Rate (CER), T3 timeouts, and T4 timeouts. For each RF signal, we included:
    - The reading for the customer at the time of the call.
    - Aggregates (mean, min, max) and the proportion of time the customer was outside of specification for each RF metric over the three days leading up to the call.

- **Modem Offline Events:** These features capture the modem's connectivity status. We included:
    - Whether the customer was currently online or offline.
    - The proportion of time the customer was offline over the past three days.

- **Gateway Health and WiFi:** These features measure the performance of the gateway and the quality of the WiFi connection within the customer premise. Specific metrics included:
    - CPU utilization.
    - Memory usage.
    - Device temperature.
    - WiFi quality of experience, which is an aggregate of Received Signal Strength Indicator (RSSI) values across the home.
    - Clear Channel Assessment (CCA).

- **Node Health and Congestion:** This feature measures the health and congestion level of the node to which the customer is connected. This gives us insight into whether network congestion might be contributing to the customer's issues.

4.2. Model Development and Training

Using the comprehensive training dataset described in the feature engineering section, we developed a classification model to predict whether a customer trouble call will require a service visit by a technician (i.e. whether their problem lies within their connection to the access network). Specifically, the model outputs the probability that a customer trouble call requires a service visit.

We experimented with various model architectures and hyperparameters to identify the best-performing model. The models and their respective hyperparameters included:

- **Logistic Regression:**
    - Regularization: Experimented with both L1 and L2 regularization techniques.
    - Regularization Strength (C): Tested various values to find the optimal trade-off between bias and variance.

- **K-Nearest Neighbors (KNN) Classifier:**
    - Number of Neighbors (num_neighbors): Experimented with different values to determine the optimal number of neighbors for classification.

- **XGBoost:**
    - Number of Estimators (num_estimators): Tested different numbers of boosting rounds to find the optimal count.
    - Maximum Depth (max_depth): Varied the maximum depth of the trees to balance model complexity and performance.
    - Learning Rate (learning_rate): Adjusted the learning rate to control the contribution of each tree.

- **Random Forest:**
    - Number of Estimators (n_estimators): Experimented with different numbers of trees in the forest.
    - Maximum Depth (max_depth): Varied the maximum depth of the trees to find the right balance between overfitting and underfitting.

After extensive experimentation and cross-validation, we selected the XGBoost model as the best performer. The selected hyperparameters for the XGBoost model were as follows:

- **Number of Estimators (n_estimators):** 100
- **Maximum Depth (max_depth):** 5
- **Random State (random_state):** 0
- **Learning Rate (learning_rate):** 0.1

The XGBoost model with these hyperparameters provided the best balance of accuracy, precision, and recall, making it the most effective model for predicting whether the root cause of a customer trouble call is inside or outside the home.

The system we built to prototype, evaluate, and deploy machine learning models is summarized in Figure 1 below.

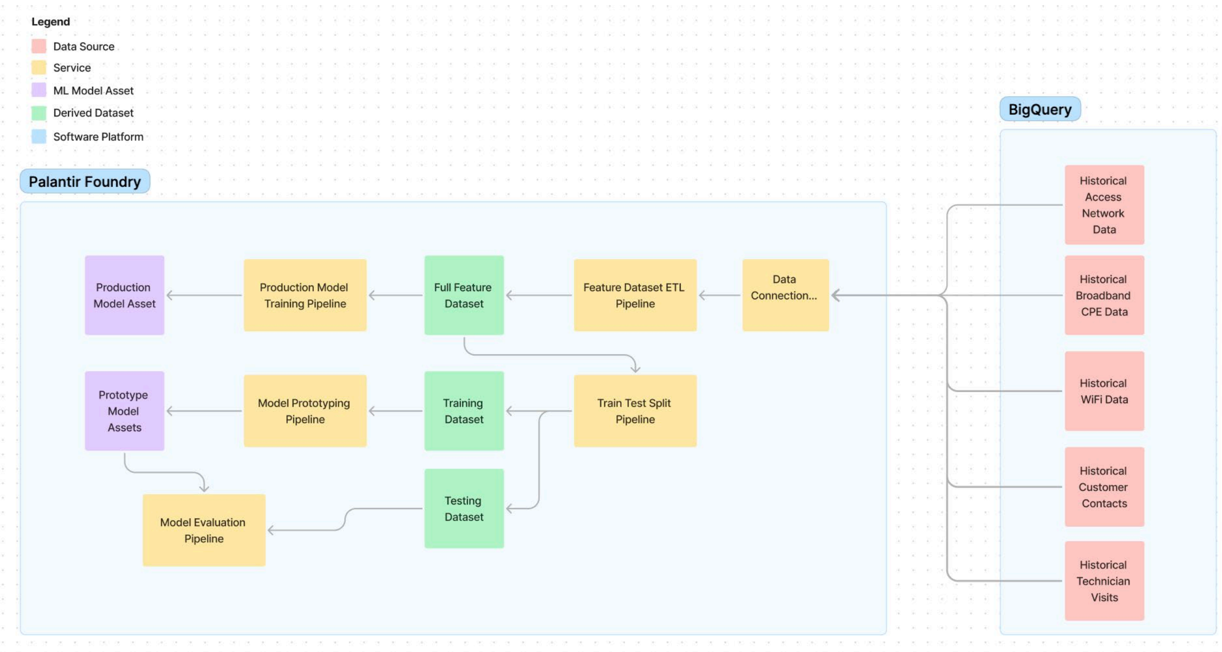

Figure 1 – Model Development Pipeline

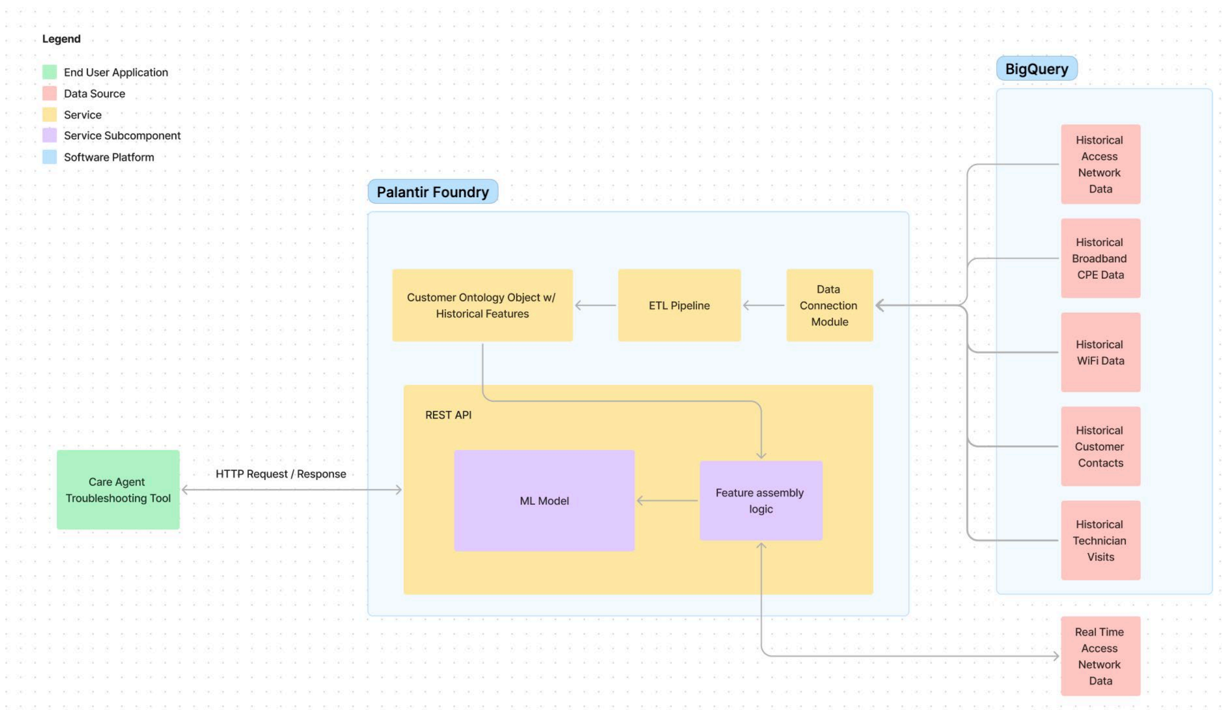

Figure 2 – Model Production Deployment System Architecture

4.3 Model Deployment

The model was exposed as a REST API service, which was built within and deployed via Foundry. To support the production deployment, we built an ETL pipeline that ingests data sources relevant to the model (e.g. RF metrics, offline events, etc) every 4 hours, and computes the necessary historical features / aggregations to be used by the model. Upon request, the REST API will:

- Be passed in a set of real time features
- Combine those features with the historical features computed in the ETL pipeline
- Invoke the model with using all of the features
- Return the classification result

The high-level technical architecture of the production model deployment is shown in Figure 2.

Today, the production model REST API has been integrated with the troubleshooting tool that the care agents use to resolve customer issues. When a customer calls and is transferred to the care agent, the troubleshooting tool makes a request to the REST API and receives a prediction result. Based on the prediction result, it will either recommend a set of modified troubleshooting steps if we are highly confident that no service visit is required or recommend the default troubleshooting steps. This is summarized in Figure 3.

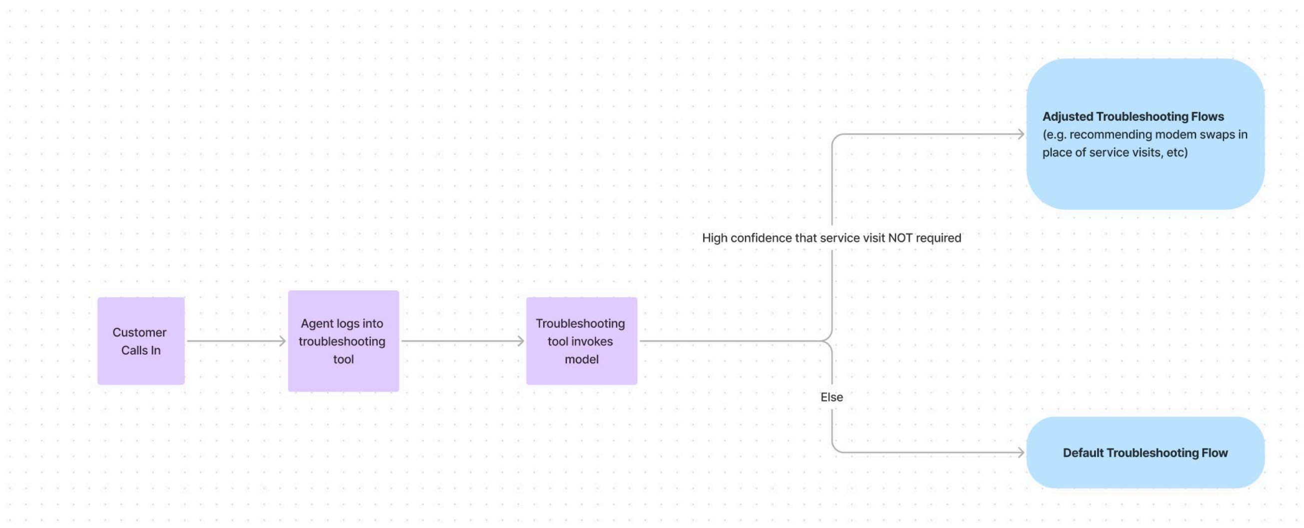

Figure 3 – Model Deployment User Journey

5. Results

The first pilot involving 75 agents has yielded promising results. The initial launch constituted a non-blocking suggestion to care agents to not send technician when the model predicted that the issue could be resolved remotely. Over the course of one month, agents from two customer care centers were observed. After the ML flow was integrated into their workflows, their performance was compared against a control group of agents from the same care centers who did not have access to the model. This integration resulted in an approximate 7% reduction of preventable truck rolls compared to the control group. This impact was achieved despite agents circumventing the recommendation by the model more than 70% of the time. This is something that will be addressed in future implementations that strictly enforce compliance to the recommendation.

To better understand the overall impact on call metrics, we monitored Average Handle Time (AHT), 7 day repeat rates (a proxy for first call resolution), Net Promoter Scores (NPS), and overall satisfaction (OSAT). With a 7% reduction in unnecessary service visits, we observed less than a 1% increase in 7-day call repeats, indicating minimal impact on first call resolution. Additionally, there was a notable decrease in AHT by 8% reflecting more efficient call handling due to the ML flow. In terms of customer satisfaction metrics, we didn't see a significant change in the values of NPS and OSAT for the customers that called agents that were part of the experiment.

With planned enhancements, including further refinements to the ML models, a broader selection of next best action flows, and expansion into additional call types, there is the possibility of increasing these savings to the mid-eight-figure range. This indicates significant positive impact on operational efficiency and cost reduction, with room for further growth as the program is scaled.

6. Future Enhancements

6.1. Overview

Future enhancements can be made to the system to both further reduce the rate at which preventable service visits are conducted and help translate the learning from customer issues into meaningful product and customer experience improvements. Specifically, we aim to:

- **"Next Best Action"**: Implement deterministic logic to give agents optimal troubleshooting recommendations based on telemetry
- **"Co-pilot"**: Use Gen AI to recommend probing questions and knowledge articles based on the agent's conversation with the customer
- **"Gen AI Insight Extraction"**: Extract insights from customer trouble calls using Gen AI

6.2. Gen AI Background

6.2.1. Overview

Generative AI enables enhanced insights into customer calls and chat transcripts, offering a more comprehensive understanding compared to traditional call listening sessions or focus groups. This technology enables the analysis of large volumes of historical troubleshooting data after scrubbing it of Personally Identifiable Information (PII) or other sensitive information. By leveraging generative AI, Altice USA can derive actionable insights, uncover hidden patterns, and improve customer service operations.

Generative AI, particularly Large Language Models (LLMs), is transforming the telecom sector by automating the analysis of vast amounts of unstructured data, which was previously analyzed manually. Traditional methods, such as call listening sessions or focus groups, are labor-intensive, time-consuming, and limited in scope. LLMs, on the other hand, can process and analyze large volumes of data in real-time, providing more comprehensive and actionable insights.

One of the primary advantages of LLMs is their scalability. These models can analyze data at an unprecedented scale, allowing telecom companies to process thousands of customer interactions simultaneously. This scalability is crucial for large organizations like Altice USA, which deal with high volumes of customer data daily. The ability to handle such large datasets ensures that no valuable information is overlooked, providing a more holistic understanding of customer issues and behaviors.

In terms of accuracy, LLMs provide a higher level of precision in identifying patterns and trends compared to manual methods. They can detect nuances in customer language and sentiment that human analysts might miss, leading to more precise insights. For example, LLMs can distinguish between different types of customer dissatisfaction, such as frustration due to long wait times or confusion over billing issues. This granular level of understanding helps telecom companies tailor their responses and solutions more effectively, while still capturing agent input through human-in-the-loop feedback.

Another significant benefit of using generative AI for data analysis is efficiency, as automation of data analysis through LLMs significantly reduces the time and effort required to derive insights. This allows telecom companies to respond to issues more quickly and implement improvements faster. Instead of spending hours or days manually reviewing call transcripts, analysts can focus on interpreting AI-generated insights and developing strategies based on those insights. This shift in focus increases overall productivity and allows for more strategic decision-making.

Cost-effectiveness is also a key advantage of automating the analysis process with generative AI. Automating the analysis process reduces the need for extensive human resources dedicated to manual data review, leading to cost savings. By reducing the reliance on human analysts for routine data processing tasks, companies can allocate their resources more efficiently. These cost savings can then be reinvested into other areas, such as improving customer service infrastructure or developing new AI capabilities.

Moreover, generative AI allows for real-time analysis and insights. Traditional methods often involve a delay between data collection and analysis. LLMs can process data as it is collected, providing immediate insights that can be acted upon quickly. This real-time capability is particularly valuable in the time-sensitive telecom industry, where customer issues and trends can change rapidly and SLAs necessitate a prompt response.

Generative AI, and specifically LLMs, provide a transformative approach to analyzing customer calls and chat transcripts in the telecom sector. The scalability, accuracy, efficiency, and cost-effectiveness of these models far surpass traditional manual methods. By leveraging these advanced technologies, Altice USA can derive actionable insights, uncover hidden patterns, and significantly improve customer service operations. The shift from manual analysis to AI-driven insights represents a substantial advancement in how telecom companies understand and respond to their customers' needs.

6.2.2. Technologies Leveraged

Large Language Models (LLMs): LLMs were pivotal in processing and analyzing textual data from calls and chat transcripts. These advanced models enabled the extraction of key phrases, sentiment analysis, and entity recognition, providing a deeper and more holistic understanding of customer interactions. By parsing through vast amounts of unstructured text data, LLMs could identify specific topics, frequently mentioned issues, and even the context surrounding customer complaints or inquiries. This capability allowed Altice USA to gain insights that were previously difficult to extract through manual analysis.

Sentiment Analysis: Leveraging LLMs, sentiment analysis was applied to customer interactions to automatically detect and categorize the sentiment expressed in the text as positive, negative, or neutral. This automated sentiment detection provided nuanced insights into customer emotions and satisfaction levels. By understanding the sentiment behind customer interactions, Altice USA could identify areas needing improvement more effectively. For example, persistent negative sentiment around a particular service feature could prompt a focused review and subsequent enhancement of that feature. Additionally, sentiment analysis enabled personalized customer service by tailoring responses based on the detected sentiment, thereby improving the overall customer experience.

Process Mining: LLMs were applied to analyze, evaluate, and extract insights from process-driven troubleshooting tools. This process is used to identify bottlenecks in existing troubleshooting flows and generate improved flows that lead to better and more accurate outcomes. For Altice USA, this technique was used by parsing historical agent troubleshooting actions, mapping them to outcomes (eg., modem reboot, equipment swap, technician dispatch) alongside existing troubleshooting rules and feeding this bulk set of datapoints into an LLM. The LLM then analyzes the agent behavioral patterns in bulk, highlighting effective and ineffective flows and patterns based on historical outcomes and suggesting adjustments to existing flows to make them more effective.

Cluster Analysis: Unsupervised machine learning techniques, such as K-means clustering and hierarchical clustering, were utilized to group similar customer issues. These clustering techniques allowed for identifying common themes and patterns in customer interactions not previously apparent through conventional methods. Grouping similar issues allowed Altice USA to prioritize and address the most frequent or impactful problems. Cluster analysis also facilitated root cause analysis by highlighting recurrent issues, enabling proactive measures to prevent future occurrences of similar problems.

Data Visualization Tools: Clusters, flows, and impacts were visualized within the Palantir platform to provide an intuitive understanding of the data. These visualizations enabled quick identification of major issues and trends. The visual interface allowed stakeholders to interact with the data dynamically, exploring various dimensions and drilling down into specific details for deeper insights. For instance, decision-makers could visualize the geographic distribution of customer complaints or track the resolution times for different types of issues. Exploring the data visually made it easier to communicate findings and collaborate on solutions across teams.

6.3. Description of Enhancements

6.3.1. *Next Best Action*

Alongside the ML models, we will continue to develop and implement additional deterministic next best action flows to guide agents through optimized troubleshooting steps. To develop these flows, we started by ingesting additional telemetry – this included telemetry related to Altice's video services (e.g. error logs, cable box health diagnostics, video QAM metrics), and additional telemetry on Altice USA's broadband CPE (e.g. the health of Wi-Fi drivers and other critical processes at the time of a call). We are in the process of building troubleshooting recommendation rules that are tailored to individual customer setups (i.e. what devices they had, and what services they were using), based on these diagnostics and defined thresholds.

Pairing ML with deterministic techniques has already shown early improvements in fiber and HFC (Hybrid Fiber-Coaxial) flows. Initial back testing indicates a significant financial impact. These techniques will be further integrated to ensure robust and reliable diagnostics, combining the predictive power of AI with the precision of deterministic methods. This hybrid approach will enable us to achieve a higher level of diagnostic accuracy, reducing the likelihood of false positives and negatives and improving overall service quality.

6.3.2. *Co-pilot*

We believe that a diagnostic driven approach to troubleshooting is the most effective way to resolve customer calls related to true service issues. However, we've found that a sizable number of customer calls are resolved through "customer education" – in other words, they are not related to true issues from a services perspective, but may be related to improper customer configuration, or simply a misunderstanding of how Altice USA's products work. In these cases, a diagnostic driven approach would not be fully effective, as telemetry would suggest that nothing is wrong with these customers.

To address these scenarios, we aim to leverage the capabilities of LLM powered agents. Specifically, we could deploy an LLM powered agent that would listen to a conversation between an agent and a customer in real time. By leveraging internal Altice USA documentation, context on Altice USA's product offering, and the customer's telemetry, the agent could guide the agent to call resolution by recommending additional probing questions and relevant documentation that could be used to educate the customer.

6.3.3. Gen AI Insight Extraction

It is difficult for Altice USA to apply learnings from customer service calls to improve their support processes and product offerings. One of the main reasons for this is that the data captured about customer calls is not accurate or high fidelity enough to be usable. Currently, customer service agents disposition calls by selecting from a dropdown of options. As such, we are not able to guarantee accuracy of these dispositions (i.e. agents may not comply), and the dispositions cannot be specific or evolve as the nature of customer issues / Altice USA's product evolves.

To address this limitation, we aim to leverage LLMs and other NLP techniques. We plan to build an LLM agent that extracts specific information about customer calls based on their transcripts. This information would include: the customer's stated issue, the agent's resolution (or lack thereof), the customer's sentiment, etc. We could then apply NLP techniques to cluster the LLM extracted fields (e.g. the customer's stated issue), which would enable Altice USA to have a specific and ever evolving grouping of customer issues. This data product could then be served to Altice USA's product and customer care teams, and enable them to derive insights such as – what products / services are our customer service agents currently having the most difficulty supporting, what product bugs / limitations are currently driving the highest number of customer calls, etc.

7. Conclusion

By leveraging key data elements from our network and in-home devices and incorporating them into our AI models we were able to operationalize a recommendation to our care agents on whether a service truck was required to address the issue experienced by a customer. In doing so we were able to show a reduction of 7% of preventable truck rolls and 8% in average handle time for our agents. This was across a sample test of 75 agents.

Enhanced granularity of data opens the door for creating more sophisticated ML models that can provide more accurate and timely diagnostics. These models will leverage both real-time data and historical trends to identify observed failures and recommend more prescriptive actions. The integration of additional ML models will further refine our understanding of network performance and customer issues, leading to more effective troubleshooting and service optimization.

Alongside these models, we will continue to develop and implement additional deterministic next best action flows to guide agents through optimized troubleshooting steps. These flows will be tailored to specific customer scenarios, ensuring that agents have the most relevant and effective recommendations at their fingertips. Pairing ML and generative AI with deterministic techniques has already shown early improvements in fiber and HFC (Hybrid Fiber-Coaxial) flows. Initial back testing indicates a significant financial impact. These techniques will be further integrated to ensure robust and reliable diagnostics, combining the predictive power of AI with the precision of deterministic methods. This hybrid approach will enable us to achieve a higher level of diagnostic accuracy, reducing the likelihood of false positives and negatives and improving overall service quality.

With increased confidence in the recommendations generated by our models and the next best action flows, we will be able to provide clear guidance to agents that lead to first call resolutions and enforce stricter agent compliance. Ensuring that agents adhere to these optimized protocols will be paramount in aligning agent performance with organizational goals. Enhanced training and a more constrained agent experience will be put in place to support agents in following the recommended actions.